心理科学进展　2022, Vol. 30, No. 8, 1682–1691
Advances in Psychological Science

https://doi.org/10.3724/SP.J.1042.2022.01682

# 新世纪 20 年国内测验信度研究*

温忠麟[1]　陈虹熹[1]　方　杰[2]　叶宝娟[3]　蔡保贞[1]

([1]华南师范大学心理学院/心理应用研究中心, 广州 510631) ([2]广东财经大学新发展研究院/应用心理学系, 广州 510320) ([3]江西师范大学心理学院/心理健康教育研究中心, 南昌 330022)

**摘　要**　随着验证性因子分析模型的应用, 信度研究进入了崭新的发展阶段。新世纪前 20 年国内有关测验信度的研究有三条发展主线。一是基于验证性因子模型的信度发展, 包括同质性系数、合成信度、最大信度等; 二是数据类型的拓展, 包括两水平和追踪数据的信度; 三是信度用途的拓展, 如评分者信度、编码者信度等。对于通常的测验(题目之间的测量误差不相关), 如果 α 系数够高, 信度就够高; 否则使用合成信度。如果一个统计模型中所有变量的合成信度都很高(超过 0.95), 使用显变量建模与使用潜变量建模的结果差别不大; 否则, 使用潜变量建模较好。

**关键词**　信度, α 系数, 同质性系数, 合成信度, 区间估计
**分类号**　B841

在心理、教育、管理等领域, 研究者广泛使用问卷测验进行实证研究, 测验信度(reliability)是测验结果的稳定性(stability)或一致性(consistency)程度, 是衡量测验质量的一个重要指标。信度反映了测验的可靠性和精确性, 即使一个完美的研究设计也无法弥补不可靠和不精确测量所带来的缺陷, 所以, 评价测验信度是进行数据分析的必要前提和重要步骤(叶宝娟 等, 2012)。

信度的定义以经典测验理论的真分数模型 $X=T+E$ 为基础, 其中 $X$ 为观测分数, $T$ 为真分数, $E$ 为测量误差。对于被试总体, 假设 $X$、$T$、$E$ 满足: 误差的均值为 0, 误差与真分数零相关, 各题目误差之间零相关。测验信度 $\rho_X$ 定义为真分数的方差与观测分数的方差之比: $\rho_X = S_T^2/S_X^2$ (Lord & Novick, 1968; 孟庆茂, 刘红云, 2002)。在有了样本数据后, 可以得到观测分数的方差, 但在经典测验理论中, 真分数的方差却无法估计, 因而研究者只好用一些替代的方法去评估信度, 这就有了人们熟知的重测信度、复本信度、分半信度、α 系数(coefficient alpha)等。

直至上世纪末, 国内信度研究的成果主要是针对 α 系数的不足提出了改进的信度估计的 β 和 γ 系数(陈希镇, 1991; 谢小庆, 1998), 但这些工作都和 α 系数一样没有从信度的定义出发, 因此提出的信度估计方法都只是某种程度上比 α 系数有改进, 但难有根本的突破。随着验证性因子分析(confirmatory factor analysis, CFA)的引入, 新世纪伊始, 国内信度研究进入了崭新的发展阶段。

新世纪前 20 年, 测验信度是仅次于结构方程模型的心理统计方法研究热点(温忠麟 等, 2021)。国内学者对信度的研究主要集中于寻找更加合适的信度指标, 以及如何在不同的情况下更加精确地估计信度。以中国知网(https://www.cnki.net/)全文数据库为数据源, 出版年限设为 2001~2020 年, 关键词包括: 信度、测验信度、重测信度、复本信度、分半信度、α 系数、同质性系数、内部一致性系数、合成信度、最大信度、评分者信度、编码者信度、信度概化, 经筛查得到有关信度的方法学研究论文 51 篇(见表 1)。从发表刊物看, 大多数文章都发表在心理学期刊上(33 篇), 这可能与心理学研究常需要使用问卷并报告问卷的信度有关, 其中《心理科学》17 篇, 《心理学探新》6 篇, 《心理学报》4 篇, 《中国临床心理学杂志》3 篇, 《心理科学进展》、《心理发展与教育》与《应用心理学》各 1 篇。此外, 《教育测

收稿日期: 2021-12-29
* 国家自然科学基金项目(32171091)资助。
通讯作者: 温忠麟, E-mail: wenzl@scnu.edu.cn

**表 1　2001~2020 年国内信度的方法学研究文献一览**

| 类别 | 文献 |
|---|---|
| α 系数 | 安胜利等(2001); 孟庆茂等(2002); 陈炳为等(2005); 席仲恩等(2007); 焦璨等(2008); 刘红云(2008); 关守义(2009); 蒋小花等(2010); 刘拓等(2011); 温忠麟等(2011); 李春会等(2012); 叶宝娟, 温忠麟(2013a); 王孟成等(2014) |
| 同质性系数 | 丁树良等(2002); 孟庆茂等(2002); 顾海根等(2005); 刘红云(2008); 陈希镇等(2011); 温忠麟等(2011, 2018); 叶宝娟, 温忠麟(2012b); 顾红磊等(2014, 2017) |
| 合成信度 | 张力为(2002); 屠金路等(2005, 2010); 徐万里(2008); 温忠麟等(2011); 叶宝娟, 温忠麟(2011, 2012a); 叶宝娟等(2013, 2014, 2015); 吴瑞林等(2012); 叶宝娟(2012); 杨强等(2014a, 2014b); 韦嘉等(2017) |
| 最大信度 | 叶宝娟, 杨强(2011); 田雪垠等(2019) |
| 单指标信度 | 方敏(2009); 王孟成等(2014) |
| 整个题目集分数的信度 | 叶宝娟, 杨强(2011) |
| 两水平研究的信度 | 叶宝娟, 温忠麟(2013b); 刘霖芯等, (2018); 田雪垠等(2019) |
| 追踪研究的信度 | 叶宝娟等(2012) |
| 评分者信度 | 严芳等(2002); 孙晓敏等(2005); 何佳等(2007); 蒋小花等(2010); 李斌等(2011) |
| 编码者信度 | 徐建平等(2005) |
| 认知诊断属性分类一致性信度 | 郭磊等(2018); 汪文义等(2018, 2020) |
| 差异分数的信度 | 关丹丹等(2005) |
| 信度概化 | 关丹丹等(2004); 焦璨等(2009) |

注：表中文献按发表时间先后排序

量与评价(理论版)》3 篇, 《中国卫生统计》3 篇, 《统计与信息论坛》2 篇, 其余 10 篇。从研究内容来看, 研究最多的是 α 系数; 其次是合成信度(composite reliability)和同质性系数(homogeneity coefficient)。

国内测验信度的研究有三条发展主线, 第一条主线是基于验证性因子模型的测验信度的发展, 从围绕 α 系数的研究发展到基于验证性因子模型的信度研究, 包括同质性系数、合成信度、最大信度(maximum reliability)、单指标信度和整个题目集分数的信度; 第二条主线是数据类型的拓展, 从单水平数据的测验信度发展到多水平数据和追踪数据(追踪数据也可看成是多水平数据)的测验信度; 第三条主线是信度用途的拓展, 从测验本身的信度发展到其他用途的信度, 如评分者信度、编码者信度、认知诊断属性分类一致性信度和差异分数的信度等。以下将按照这三条主线逐一评述国内新世纪前 20 年的信度研究。

## 1　有关 α 系数的研究

### 1.1　α 系数的点估计和区间估计

α 系数是最常用的信度指标, 信度的发展大都以 α 系数为基础, α 系数的计算公式为:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\text{各题方差之和}}{\text{总分方差}}\right) \tag{1}$$

其中 $k$ 为量表中的题数, 根据样本计算 α 系数时将方差改为样本方差便可。α 系数可以用 SPSS 软件计算得出, 也可在 SAS 软件中调用 PROC CORR 过程计算 α 系数, 只要在选择项加上参数 α 即可(马文军, 潘波, 2000), 也可用 Mplus 软件编写程序计算 α 系数(王孟成, 叶宝娟, 2014)。

α 系数是一个总体参数, 在实证研究中通常用样本的 α 系数来估计, 最好同时计算其置信区间, 以此得到在所研究的总体上重复取样时 α 系数的精确性(Raykov & Shrout, 2002; Zinbarg et al., 2006)。尤其在样本容量不大、α 较小时, 非常有必要报告 α 系数的置信区间(Maydeu-Olivares et al., 2007)。叶宝娟和温忠麟(2013a)介绍了 10 种计算 α 系数置信区间的方法, 并通过模拟研究比较了其中较好的 7 种方法(包括 Fisher 法、Bonett-02 法、Bonett-10 法、精确 Koning-Franses 法、渐近 ID 法、渐近 Koning-Franses 法和 ADF 法)。结果发现 Bonett-10 法和精确 Koning-Franses 法较好。这两种方法都比较简单, 只需要样本的 α 值、测验题数、被试人数及 $F$ 临界值, 通过简单的运算便可得到 α 系数的置信区间。

### 1.2 α 系数和信度的关系

许多研究表明, α 系数不能很好地估计测验信度(陈炳为 等, 2005; 刘拓, 戴晓阳, 2011; 李春会, 朱永忠, 2012)。刘红云(2008)通过模拟研究表明, 在基本 τ-等价(essentially τ equivalent)测验的条件下(即任意两个题目的真分数只相差一个常数, Graham, 2006), α 系数于测验信度, 否则 α 系数容易低估测验信度。有时候 α 系数甚至还会出现负值(席仲恩, 汪顺玉, 2007)。温忠麟和叶宝娟(2011)通过梳理文献后指出, 如果(i)各题的误差不相关(这个条件容易满足); (ii)测验是基本 τ 等价(这个条件很强, 通常的测验难以满足), α 系数等于测验信度; 如果满足条件(i), 但不满足条件(ii), α 系数小于信度。总之, 如果各题的误差不相关, α 系数是信度的下限(即有可能低估信度); 否则 α 系数有可能高估信度。多数情况下, 各题的误差是不相关的, 若 α 系数高到可以接受, 那么测验信度就可以接受, 所以 α 系数还可以继续使用(温忠麟, 叶宝娟, 2011)。

### 1.3 对 α 系数的误解和误用

传统上将 α 系数称为内部一致性信度或者同质性系数, 但实际上 α 系数不能用来衡量测验的内部一致性, 也不能用来衡量测验的同质性(温忠麟, 叶宝娟, 2011), 因为题目数量的增加会导致 α 系数的增加, 哪怕是多维度的测验, 只要题目够多, α 系数就会高(孟庆茂, 刘红云, 2002)。已有研究发现 α 系数高不代表测验是同质的(刘红云, 2008)。为了避免研究者为提高 α 系数而增加多余条目的行为, 有人认为 α 系数不宜超过 0.9 (安胜利, 陈平雁, 2001; 孟庆茂, 刘红云, 2002)。后面我们会看到, 合成信度可以用来衡量测验的内部一致性, 同质性则要使用同质性系数来衡量(见第 2 节)。

在应用 α 系数的过程中容易忽视正态分布这一使用前提。因为 α 系数以经典测量理论的真分数模型为基础, 但经典测量理论是以正态分布为前提的。焦璨等(2008)通过模拟研究表明, 当测验数据为非正态时, 偏度越大, α 系数越小。焦璨等建议, 先将非正态数据进行聚类, 假设聚为 3 个组, 则分别求 3 个组的 α 系数, 用多个 α 系数来描述测验可靠性。关守义(2009)进一步指出, α 系数在实际应用中除了需要满足正态分布的前提, 还需要确保每个个体方差相同、每次观测均相互独立, 并且各测量题目需要具有相同的心理刻度。

## 2 基于验证性因子模型的测验信度

随着验证性因子模型(包括双因子模型)的引入, 信度研究得到了迅猛发展, 其中研究最多且成果最丰富的当属同质性系数和合成信度。

### 2.1 同质性系数

同质性是指所有题目都测量了相同的特质(Revelle & Zinbarg, 2009; 刘红云, 2008), 如果所有题目之间的相关都高, 则同质性高。无论单维还是多维测验, 都可以考虑测验同质性。

#### 2.1.1 单维测验的同质性系数

其实新世纪前后国内已经有学者发现 α 系数不能很好地衡量同质性, 提出了一些新的同质性指标。陈希镇(1991)提出了 β 系数, 谢小庆(1998)提出了 γ 系数, 丁树良和周新莲(2002)提出 ξ 系数。这些系数都只是某种程度上比 α 系数有改进, 但也和 α 系数一样, 没有从信度的定义出发, 所以没有根本上的突破。

基于验证性因子模型, Raykov (2001)提出用 ρ 系数作为单维测验(也称为同属测验)的同质性系数, 这是方法上的突破。在建立单因子模型后, 整份测验的总分 = (题目的因子负荷之和) × 因子 + (题目误差之和), 加号前面的为真分数部分, 加号后面的为误差部分。这样, 总分的方差就可以分解为真分数方差和误差方差。将信度的定义应用于总分, 就得到 ρ 系数, 它等于测验总分的方差中, 真分数方差所占的比例(Raykov, 2001; 陈希镇, 李学娟, 2011)。ρ 系数可以用任意一款结构方程软件计算得到。

顾海根和李超(2005)采用概化理论的研究方法, 对 ρ 系数、α 系数、β 系数、γ 系数、ξ 系数进行了比较研究。结果发现, ρ 系数最优, 表现在 ρ 系数最接近信度的真值, α 系数最劣, β、γ、ξ 系数基本处于一个水平, 介于 ρ 和 α 系数之间。因此, 应当将 ρ 系数作为单维测验的同质性系数指标。

#### 2.1.2 多维测验的同质性系数

对于多维测验, 在决定将多个维度的测验分数合成测验总分时, 应当考虑测验同质性的高低。如果测验同质性高(例如大于 0.5), 合成总分是有意义的(温忠麟 等, 2018); 如果同质性太低, 合成总分没有什么意义, 以合成总分为基础进行的统计分析也就没有什么意义, 这时应当以维度为变量进行统计分析。

估计同质性系数可以使用双因子模型(bi-factor model, 详见：顾红磊 等, 2014)。在双因子模型中, 测验总分的方差就可以分解为三部分：全局因子分数的方差、局部因子分数的方差和误差方差。测验的同质性系数定义为：测验总分的方差中, 全局因子分数方差所占的比例, 有些文献将其记为 $\omega_h$ (Revelle & Zinbarg, 2009; 温忠麟, 叶宝娟, 2011)。

叶宝娟和温忠麟(2012b)用 Delta 法推导出计算同质性系数的标准误公式, 进而计算其置信区间。他们通过模拟比较了用 Delta 法和 Bootstrap 法计算的置信区间, 发现两者差异很小。他们还提供了简单的计算多维测验的同质性系数及其置信区间的 LISREL 和 Mplus 程序。

与同质性密切相关的一个概念是单维性。在双因子模型中, 将全部题目的全局因子分数的方差相加是全局因子解释的方差, 将全部题目的局部因子分数的方差相加后再加上全局因子解释的方差就是公共方差。全局因子对公共方差的解释比例(explained common variance, ECV) = (各题的全局因子分数的方差之和) / (各题的全局因子分数的方差之和 + 各题的局部因子分数的方差之和)。ECV 通常作为单维性指标(Bentler, 2009), 用来判断多维测验的单维倾向性的程度。如果 ECV 超过 0.7, 可以认为测验是单维的(顾红磊, 温忠麟, 2017; Reise, 2012)。ECV 指标可用 Mplus 软件进行计算(王孟成, 叶宝娟, 2014; 顾红磊, 温忠麟, 2017)。

综上可知, 同质性系数和单维性指标 ECV 是两个同源指标, 都源于双因子模型将每个题目分解为三个部分, 如果从整份测验的总分入手进行分析, 则可得同质性系数; 如果从题目的方差入手进行分析, 则可得 ECV。随着全局因子的方差的变大, 同质性系数和 ECV 都会变高。两者的区别也明显, 因为 ECV 没有涉及误差方差, 单维测验的同质性不一定高(因为可能误差方差大)。但同质性系数越高, ECV 也越高。

#### 2.1.3　题目表述效应对同质性系数的影响

题目表述效应是指由题目表述方式的差异(如正向题和反向题)引起的与测量内容无关的系统变异。题目表述效应模型本质上是一种双因子模型, 包括全局因子(所测特质 $G$, 影响全部题目)、局部因子(如正向题目效应因子 $F1$, 反向题目效应因子 $F2$)和测量误差。评价这类测验的同质性系数可以了解, 在排除了题目表述效应和测验误差引起的变异之后, 所测特质的变异占总变异的比例, 进而评价合成总分是否有意义。顾红磊和温忠麟(2014)发现忽视题目表述效应会高估测验的同质性系数。韦嘉等(2017)发现忽视题目表述效应, 还会高估测验的 α 系数和合成信度。

### 2.2　合成信度

#### 2.2.1　合成信度的点估计和区间估计

合成信度是量表的合成分数(均值或者总分)的信度。对于单维测验, 合成信度与同质性系数相同(温忠麟, 叶宝娟, 2011), 即测验总分的方差中真分数方差所占的比例。单维测验的合成信度可用 SPSS 软件(杨强 等, 2014b)、LISREL 和 Mplus 软件(温忠麟, 叶宝娟, 2011)计算得到。

对于多维测验, 使用双因子模型将总分的方差分解为三部分：全局因子分数的方差、局部因子分数的方差和误差方差。测验的合成信度定义为：测验总分方差中, 全局因子和所有局部因子分数方差所占的比例, 有些文献将其记为 ω (Revelle & Zinbarg, 2009; 温忠麟, 叶宝娟, 2011)。总分的方差中, 如果将误差方差之外的都理解为真分数的方差, 按信度定义计算得到的就是合成信度。多维测验的合成信度可用 LISREL (徐万里, 2008; 屠金路 等, 2010)和 Mplus (王孟成, 叶宝娟, 2014)等结构方程软件计算得到。

值得注意的是, 合成信度在计算测验总分的时候, 通常直接将题目得分相加求和, 即将测验所有的题目赋予了同样的权重(权重为 1)。也有研究者利用验证性因子分析的结果, 选择一组权重(每个题目的权重=该题目的因子负荷/该题目的误差方差), 将题目得分乘以该题的权重, 再求和合成一个总分, 此时求得的合成信度达到最大值, 称为最大信度(Fu et al., 2018; 田雪垠 等, 2019; 叶宝娟, 杨强, 2011)。最大信度即可用于通常的单维测验(叶宝娟, 杨强, 2011)和多维测验(Fu et al., 2018), 也可用于被试有层级结构的测验(即多水平测验; 田雪垠 等, 2019)。

有三种方法可以估计合成信度的标准误进而计算合成信度的置信区间：Bootstrap 法(屠金路 等, 2005)、Delta 法、直接用结构方程建模软件输出的标准误。叶宝娟和温忠麟(2011)比较了以上三种方法在计算单维测验合成信度的置信区间中的

表现，推荐用Mplus软件估计Delta法的单维测验合成信度的置信区间。后续的一系列研究都表明，无论单维还是多维，是否偏态分布，测验误差是否相关，都推荐使用 Delta 法估计合成信度的置信区间(杨强 等, 2014a; 叶宝娟, 2012; 叶宝娟, 温忠麟, 2012a; 叶宝娟, 杨强, 2014, 2015)。

### 2.2.2 合成信度与内部一致性、同质性的关系

内部一致性可以定义为题目之间的相关性(Revelle & Zinbarg, 2009)，如果同一维度内部的题目之间相关高，则内部一致性高。对于多维量表，内部一致性应当理解为同一维度内部的题目之间的相关性，而不是全部题目之间的相关性。这样，合成信度可以理解为内部一致性信度(Bentler, 2009; 温忠麟, 叶宝娟, 2011)。同质性高的测验，内部一致性也高，但反过来不一定成立(张力为, 2002)。

可以证明同质性系数不超过合成信度(因为合成信度的分子中包含局部因子方差)，当且仅当局部因子不存在时(即单维)，两者相等(温忠麟, 叶宝娟, 2011)。不论误差是否相关，合成信度都不超过测验信度(温忠麟, 叶宝娟, 2011)，即同质性系数≤合成信度≤测验信度。因此，用合成信度来估计测验信度更为准确。温忠麟和叶宝娟(2011)总结出一个测验信度分析流程(见图 1)，可以对量表合成分数的信度做出评价。

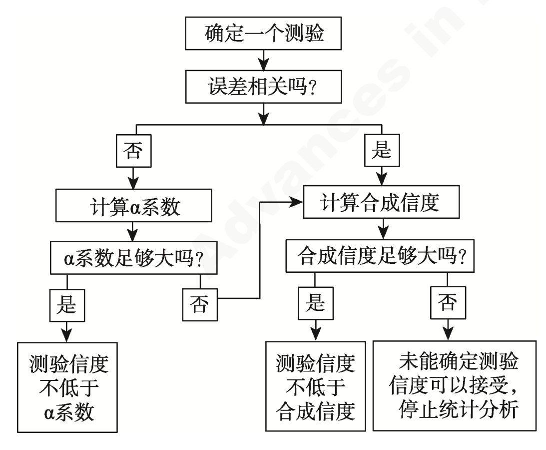



图 1 测验信度分析流程(温忠麟, 叶宝娟, 2011)

### 2.2.3 合成信度的实际意义

研究变量之间关系既有基于显变量(合成分数)的分析(可以使用回归模型)，也有基于潜变量(带有指标)的分析(可以使用结构方程模型)。什么时候使用显变量分析已经足够，什么时候需要潜变量分析才好，判断的主要依据就是量表的合成信度。两个显变量之间的相关系数，等于两个相应的潜变量之间的相关系数乘以两个合成信度的几何平均(侯杰泰 等, 2004)。如果两个合成信度都超过 0.95 (题目较多的许多量表都满足)，使用显变量分析与使用潜变量分析的结果差别不大，否则，使用潜变量分析较好。对于回归系数，关键是自变量的合成信度(因为因变量通过残差考虑了测验误差)。使用显变量得到的回归系数，等于使用潜变量得到的回归系数乘以自变量的合成信度。如果自变量的合成信度超过 0.95，使用显变量分析与使用潜变量分析的结果差别不大，否则，使用潜变量分析较好。

## 2.3 其他测验信度

### 2.3.1 单指标信度

根据真分数模型，测验中的单个题目是无法按信度的定义计算其信度的。但基于验证性因子分析，真分数的方差也是可以估计的，因而可以估计单个题目的信度，即单指标信度。单指标信度反映单一题目得分受潜变量影响的程度，其值越高，表示真分数所占的比重越大(方敏, 2009)。王孟成和叶宝娟(2014)给出了计算单指标信度的 Mplus 程序。对于完全标准化估计，题目负荷的平方就是单指标信度。

### 2.3.2 整个题目集分数的信度

用合成信度和最大信度衡量测验的信度是有前提的，即测验的各个题目可以相加得一总分。而在实际应用中，有些测验的各个题目相加得一总分并没有多大意义，虽然此时仍可计算合成信度及最大信度，但没有意义。Alonso 等(2010)用验证性因子分析推导出了两个新的信度系数 $R_T$ 和 $R_\Lambda$。$R_T$ 信度系数是用观测分数与误差分数的方差-协方差矩阵的迹，来概括观测分数与误差分数的变异得到的；而 $R_\Lambda$ 是用观测分数与误差分数方差−协方差矩阵的行列式，来概括观测分数与误差分数的变异得到的。

叶宝娟和杨强(2011)比较了合成信度、最大信度、$R_T$ 和 $R_\Lambda$，并讨论了这 4 种信度系数估计方法的差异：(1)信度计算时对每个题目分数的处理方法不一样。合成信度及最大信度是将各题目分数单位加权或不等加权合成总分，计算的是测验总分的信度，而 $R_T$ 和 $R_\Lambda$ 系数计算的是整个测验题

目集分数的信度。(2)测验长度对信度的影响不一样。随着题目的增多, 合成信度不一定变大, 如果加入质量不好的题目(如题目的因子负荷小), 合成信度和 $R_T$ 都可能降低; 而最大信度和 $R_\Lambda$ 会随着题目的增多而增大。(3)信度系数的数值大小不同。最大信度= $R_\Lambda$ ≥合成信度≥ $R_T$。

## 3　特殊数据类型的测验信度

前面介绍的信度用于常规的测验数据。对于有层级结构的数据(多水平数据)、追踪研究的重复测量数据(纵向数据), 需要有相应的方法得到更准确的信度估计。

### 3.1　两水平研究的信度

在心理、教育、管理等研究领域中, 经常会遇到两水平的数据结构, 如学生嵌套于班级中, 员工嵌套于企业中, 这样的两层数据结构能够更准确地研究变量之间的关系。叶宝娟和温忠麟(2013b)用两水平验证性因子分析模型推导出两水平研究中单维测验的信度公式, 无论组间因子负荷是自由还是固定都适用。组间因子负荷自由估计的两水平研究中, 单维测验信度的点估计可用 Mplus 软件得到(叶宝娟, 温忠麟, 2013b)。但如何得到单维测验信度的区间估计还亟待解决。

田雪垠等(2019)以两层数据为例讨论多水平研究的信度估计。将观察分数分解为层 1 真分数和层 2 真分数、层 1 误差和层 2 误差四个部分。然后分别估计层 1 信度和层 2 信度, 包括层 1 和层 2 的 α 系数、合成信度和最大信度。例如, 将 α 系数公式分别应用于层 1 的方差和层 2 的方差, 得到层 1 的 α 系数和层 2 的 α 系数。他们使用 Mplus 软件展示了如何利用两水平验证性因子分析计算两水平多维测验的信度。但如何得到多维测验信度的区间估计还亟待解决。另外, 刘霖芯等(2018)将单层数据($n$ 个被试完成 $k$ 个题目)看成是题目嵌套于被试的两层数据(题目为层 1, 被试为层 2), 利用两水平模型计算 α 系数。

### 3.2　追踪研究的信度

有研究者用体现追踪数据特点的数学模型提出相应的信度估计, 包括基于单纯形模型的 $\rho_w$, 基于概化单纯形模型的 $\rho(S_w)$, 但 $\rho_w$ 和 $\rho(S_w)$都只估计了单个时间点测验的信度, 而没有给出整个追踪研究测验的信度。还有研究者基于线性混合模型, 利用前面介绍的计算 $R_T$ 和 $R_\Lambda$ 的思想定义了追踪数据中的 $R_T$ 和 $R_\Lambda$, $R_T$ 和 $R_\Lambda$ 既可估计追踪研究中单个时间点的测验信度, 也可估计整个追踪研究的测验信度, 推荐同时使用 $R_T$ 和 $R_\Lambda$ 来估计追踪研究的测验信度(叶宝娟 等, 2012)。但是在非线性条件以及非平衡设计等条件下, 它们的适用性还有待进一步的研究。实际上, 追踪数据还可看成重复测量的时间点嵌套于被试的两水平嵌套数据, 用两水平信度测量方法进行信度估计。关于 $R_T$ 和 $R_\Lambda$ 与两水平信度系数在追踪数据的信度计算中的表现孰优孰劣, 有待进一步研究。

## 4　其他用途的测验信度

除了用来评价测验(如问卷和试题)结果的一致性外, 信度还可以有其他用途, 例如评价不同评分者对被试作答的评分的一致性(评分者信度)、评价不同编码者对相同文本独立编码的一致性(编码者信度)、评价认知诊断属性分类的一致性(认知诊断属性分类一致性信度)、评价培训或者练习效果的一致性(差异分数的信度)等。

### 4.1　评分者信度

评分者信度的计算方法有相关法(孙晓敏, 张厚粲, 2005; 何佳 等, 2007; 蒋小花 等, 2010), 百分比法(孙晓敏, 张厚粲, 2005)和基于概化理论的方法(严芳, 李伟明, 2002; 李斌 等, 2011)。相比相关法和百分比法, 概化理论对评分者一致性的估计更为灵活(所需前提假设更少, 适用面更广)和主动(不仅可以得到概化系数, 还可以根据所得到的方差分量估计值算出为达到一定的概化系数, 选择多少评分者是合适的), 孙晓敏和张厚粲(2005)推荐用概化理论估计表现性评价中的评分者信度。

### 4.2　编码者信度

检验编码者信度的方法有归类一致性指数、编码信度系数、相关系数、中位数检验、概化系数(徐建平, 张厚粲, 2005)。其中, 归类一致性指数是指对编码归类相同数占归类总数的百分比, 因此其稳定性更多地受相同编码数目的影响, 相同编码数据越多, 归类一致性指数越高; 概化系数则受编码者和编码题目数量的影响。具体地, 编码者侧面、以及与编码者相关的交互效应变异分量越小, 编码者一致性就越高。在编码题目数量较小时, 概化系数的增幅较大(徐建平, 张厚粲, 2005)。

### 4.3 认知诊断属性分类一致性信度

对于认知诊断的属性分类一致性信度的点估计，可用改进后的 α 系数法(汪文义 等, 2018)、四分相关法(郭磊，张金明，2018)、一致性法(郭磊，张金明，2018；汪文义 等，2018)、基于 Bootstrap 的积差相关法和修正的一致性法(郭磊，张金明，2018)进行估计。郭磊和张金明(2018)的模拟研究表明，积差相关法表现最优(平均偏差的绝对值更接近 0 和误差均方根指标最小)，修正的一致性法和一致法居中，四分相关法最差。

对于认知诊断属性分类一致性信度的区间估计，汪文义等(2018)在一致法获得属性分类一致性的点估计的基础上，比较了三种估计信度置信区间的方法：Bootstrap 法、平行测验配对法和平行测验法，推荐使用 Bootstrap 法估计认知诊断属性分类一致性信度的置信区间。汪文义等(2020)进一步发现，用 Bootstrap 法估计的属性分类一致性信度平均数和标准误在不同研究条件的模型选择率较稳定，总体表现较好。

### 4.4 差异分数的信度

差异分数(也称增长分数)是指同一批被试两次测试的得分之差，用以判断培训或者练习的效果。关丹丹等(2005)给出了差异分数的信度点估计的计算公式，差异分数的信度不超过两次测试中信度相对较高的那次测试的信度。两次测试的信度、标准差和相关都会影响差异分数的信度。

## 5 讨论与拓展

新世纪 20 年来，国内学者努力探索如何更准确地估计测验的信度，既包括理论层面的探索(从经典测验理论→概化理论；陈社育，余嘉元，2001)，也包括工具层面的探索(从无因子分析模型→验证性因子分析模型→双因子模型)，使得信度领域的方法学研究取得了长足的发展，加深了我们对信度的认识。本文从有关 α 系数的研究开始，系统回顾了这期间国内有关信度的研究，重心是基于验证性因子分析模型的信度，也包括两水平和追踪数据的信度、其他用途的测验信度等。但也还有一些尚未介绍的发展情况需要补充一下。

第一，国内信度研究在信度的元分析方面也有发展。信度的元分析有两类研究。一类是利用变化系数模型对单维测验的合成信度进行元分析，并提出用 Delta 法估计合成信度元分析置信区间(叶宝娟 等，2013)。他们还以区间覆盖率为衡量指标，用模拟研究证明 Delta 法的合成信度元分析区间估计的方法是得当的。另一类是信度概化(reliability generalization)，信度概化是概化理论的应用，它是以某一特定的测验工具(如问卷)在不同研究中的信度系数作为研究样本，对这些信度系数作再研究，探究影响信度的变量，即使信度系数发生变化的预测源，并研究与信度系数有关的测量条件及因素的变异性(关丹丹，张厚粲，2004；焦璨 等，2009)。

第二，已有一些学者尝试在传统心理测验中引入计算机化自适应测验技术(例如，李宇斌 等，2020；汪大勋，涂冬波，2021；张龙飞 等，2020)，但目前还没有合适的方法估计计算机化自适应测验信度，有待研究。

为了更好地对信度的当下研究有所了解，下面从三个方面介绍国外期刊信度研究的情况，或许可以发现一些值得未来探索和拓展的方向。

### 5.1 α 系数该不该放弃?

McNeish (2018)认为 α 系数过时了，建议用合成信度、最大信度等替代 α 系数。Raykov 和 Marcoulides (2019)则认为，在某些条件下，α 系数还可以作为信度的估计值，不用放弃。Sijtsma 和 Pfadt (2021)指出，即使题目误差相关，α 系数的属性仍保持不变。国外学者现在还在争论的这个问题，国内学者在 10 年前已经说得相当清楚。温忠麟和叶宝娟(2011)已明确指出，对于大多数测验来说，假设误差不相关是合理的，如果 α 系数高到可以接受，那么测验信度就可以接受。因而对于大多数测验来说，计算并报告 α 系数，已经足以支持测验信度。所以，多数情况下，α 系数还可以继续使用。

### 5.2 有关合成信度的研究

Edwards 等(2021)比较了单维的合成信度、多维的合成信度、信度的最大下限和 α 系数的表现，结果发现合成信度和 α 系数比较准确地反映了总体信度，且信度估计受样本大小、基本 τ-等价的违反程度、总体信度大小和题目数量的影响。合成信度受样本大小和题目数量的影响更大，特别是当总体信度低的时候，而 α 系数对违反 τ-等价性的程度比较敏感。

Padilla 和 Divers (2016)比较了 6 种合成信度的置信区间获得方法(不包括贝叶斯法)，结果发

现 bootstrap 置信区间表现最优。Kelley 和 Pornprasertmanit (2016)比较了 4 种信度系数的置信区间，包括类别变量的合成信度，建议使用 bootstrap 置信区间。Pfadt 等(in press)提出在贝叶斯框架下，利用 Gibbs 抽样得到信度系数的后验分布后，估计信度系数的可信区间。模拟研究表明，在无信息先验条件下，95%的贝叶斯可信区间与 95%的 bootstrap 置信区间相当。

如果因子模型有跨因子负荷却被忽略，结果会如何？Fu 等(2022)用模拟研究比较了探索性结构方程模型(麦玉娇，温忠麟，2013)和验证性因子模型在求合成信度中的表现。结果表明，基于探索性结构方程模型和验证性因子模型得到的合成信度相当接近，说明忽略跨因子负荷对合成信度的估计影响不大。

Lai 等(2020)将合成信度拓展到多水平模型中，定义了 6 种适用于多水平数据的合成信度指标，并给出 R 和 Mplus 程序计算信度的置信区间。

### 5.3　基于概化理论的信度研究

Scherer 和 Teo (2020)指出信度概化存在三个不足：信度系数估计中存在不切实际的假设(例如，α 系数的基本 τ-等价假设)；忽略量表总分和分量表分数的信度系数的相关性；不同类型的信度系数之间缺乏可比性。他们提出元分析结构方程(meta-analytic structural equation modeling, MASEM)来解决这三个不足，进行信度概化分析。ten Hove 等(in press)将评分者信度拓展到多水平模型，用概化理论考察多水平观测数据的评分者信度，用马尔可夫链蒙特卡罗法来估计多水平观测数据的方差。

## 参考文献


安胜利，陈平雁．(2001)．量表的信度及其影响因素．*中国临床心理学杂志*, *9*(4), 315–318.

陈炳为，许碧云，倪宗瓒，杨惠芳．(2005)．证实性因子分析在量表信度中的应用研究．*中国卫生统计*, *22*(4). 261–263.

陈社育，余嘉元．(2001)．经典真分数理论与概化理论信度观评析．*心理科学进展*, *9*(3), 258–263.

陈希镇．(1991)．如何正确使用信度估计公式．*心理学报*, *24*(1), 41–49.

陈希镇，李学娟．(2011)．结构方程模型下的信度估计．*统计与决策*, *27*(1), 13–15.

丁树良，周新莲．(2002)．一种新的信度估计．*江西师范大学学报(自然科学版)*, *26*(3), 222–224.

方敏．(2009)．结构方程模型下的信度检验．*中国卫生统计*, *26*(5), 524–526.

顾海根，李超．(2005)．同质信度多种指标的比较研究．*心理科学*, *28*(5), 1196–1198.

顾红磊，温忠麟．(2014)．项目表述效应对自陈量表信效度的影响——以核心自我评价量表为例．*心理科学*, *37*(5), 1245–1252.

顾红磊，温忠麟．(2017)．多维测验分数的报告与解释：基于双因子模型的视角．*心理发展与教育*, *33*(4), 504–512.

顾红磊，温忠麟，方杰．(2014)．双因子模型：多维构念测量的新视角．*心理科学*, *37*(4), 973–979.

关丹丹，张厚粲．(2004)．信度的再认识与信度概括化研究．*心理科学*, *27*(2), 445–448.

关丹丹，张厚粲，李中权．(2005)．差异分数的信度分析．*心理科学*, *28*(1), 161–163.

关守义．(2009)．克龙巴赫 α 系数研究述评．*心理科学*, *32*(3), 685–687.

郭磊，张金明．(2018)．使用 Bootstrap 方法计算认知诊断评估中的信度．*心理学探新*, *38*(5), 433–439.

何佳，何惧，席雁，徐超．(2007)．评分者信度的分析方法简介及比较．*中国现代医生*, *45*(6), 76–77.

侯杰泰，温忠麟，成子娟．(2004)．*结构方程模型及其应用*．北京：教育科学出版社．

蒋小花，沈卓之，张楠楠，廖洪秀，徐海燕．(2010)．问卷的信度和效度分析．*现代预防医学*, *37*(3), 429–431.

焦璨，吴利，张敏强，张文怡．(2009)．信度概化研究的新进展评析．*学术研究*, *52*(2), 54–59.

焦璨，张敏强，黄庆均，张文怡，黎光明．(2008)．非正态分布测量数据对克隆巴赫信度 α 系数的影响．*应用心理学*, *14*(3), 276–281.

李斌，辛涛，张淑梅，孙佳楠．(2011)．多评分者多任务情境下评分者信度的模型拟合研究．*湖南师范大学教育科学学报*, *10*(6), 107–110.

李春会，朱永忠．(2012)．基于信度系数与 α 系数分析结构方程模型．*暨南大学学报(自然科学与医学版)*, *33*(3), 250–252.

李宇斌，蔡艳，涂冬波．(2020)．手机依赖的计算机化自适应测量及其效果评估．*心理科学*, *43*(3), 748–755.

刘红云．(2008)．α 系数与测验的同质性．*心理科学*, *31*(1), 185–188.

刘霖芯，张韬，杨珉．(2018)．利用多水平模型计算及校正 Cronbach alpha 系数．*中国卫生统计*, *35*(6), 838–842.

刘拓，戴晓阳．(2011)．不拟合被试对测验信、效度的影响．*中国临床心理学杂志*, *19*(6), 743–745.

马文军，潘波．(2000)．问卷的信度和效度以及如何用 SAS 软件分析．*中国卫生统计*, *17*(6), 364–365.

麦玉娇，温忠麟．(2013)．探索性结构方程建模(ESEM)：EFA 和 CFA 的整合．*心理科学进展*, *21*(5), 934–939.

孟庆茂，刘红云．(2002)．α 系数在使用中存在的问题．*心理学探新*, *22*(3), 42–47.

孙晓敏, 张厚粲. (2005). 表现性评价中评分者信度估计方法的比较研究——从相关法、百分比法到概化理论. *心理科学, 28*(3), 646–649.

田雪垠, 郑蝉金, 郭少阳, 贺冠瑞. (2019). 基于多层验证性因素分析的各种信度系数方法. *心理学探新*, *39*(5), 461–467.

屠金路, 金瑜, 王庭照. (2005). bootstrap 法在合成分数信度区间估计中的应用. *心理科学, 28*(5), 1199–1200.

屠金路, 王庭照, 金瑜. (2010). 结构方程模型下多因子非同质测量合成分数的信度估计. *心理科学, 33*(3), 666–669.

汪大勋, 涂冬波. (2021). 认知诊断计算机化自适应测量技术在心理障碍诊断与评估中的应用. *江西师范大学学报(自然科学版), 45*(2), 111–117.

王孟成, 叶宝娟. (2014). 通过 Mplus 计算几种常用的测验信度. *心理学探新, 34*(1), 48–52.

汪文义, 方小婷, 叶宝娟. (2018). 认知诊断属性分类一致性信度区间估计三种方法. *心理科学, 41*(6), 1492–1499.

汪文义, 朱黎君, 叶宝娟, 方小婷. (2020). Bootstrap 区间估计在认知诊断模型误设中的应用. *心理科学*, *43*(6), 1498–1505.

韦嘉, 郭磊, 张进辅. (2017). 表述效应对平衡量表内部一致性信度的影响. *西南大学学报(自然科学版)*, *39*(8), 133–139.

温忠麟, 方杰, 沈嘉琦, 谭倚天, 李定欣, 马益铭. (2021). 新世纪 20 年国内心理统计方法研究回顾. *心理科学进展, 29*(8). 1331–1344.

温忠麟, 黄彬彬, 汤丹丹. (2018). 问卷数据建模前传. *心理科学, 41*(1), 204–210.

温忠麟, 叶宝娟. (2011). 测验信度估计: 从 α 系数到内部一致性信度. *心理学报, 43*(7), 821–829.

吴瑞林, 袁克海. (2012). 基于结构方程模型的合成信度及其使用问题研究. *统计与信息论坛, 27*(12), 14–20.

席仲恩, 汪顺玉. (2007). 论负克伦巴赫 alpha 系数和分半信度系数. *重庆邮电大学学报(自然科学版), 19*(6), 785–787.

谢小庆. (1998). 信度估计的 γ 系数. *心理学报*, *30*(2), 193–196.

徐建平, 张厚粲. (2005). 质性研究中编码者信度的多种方法考察. *心理科学, 28*(6), 152–154.

徐万里. (2008). 结构方程模式在信度检验中的应用. *统计与信息论坛, 23*(7), 9–13.

严芳, 李伟明. (2002). 用结构方程建模(SEM)估计概化理论(GT)中的评分者信度. *心理学报, 34*(5), 534–539.

杨强, 叶宝娟, 温忠麟. (2014a). 两种估计多维测验合成信度置信区间方法比较. *心理学探新, 34*(1), 43–47.

杨强, 叶宝娟, 温忠麟. (2014b). 用 SPSS 软件计算单维测验的合成信度. *中国临床心理学杂志, 22*(3), 496–498.

叶宝娟. (2012). 偏态分布下单维测验合成信度三种区间估计的比较. *教育测量与评价(理论版), 5*(10), 28–32.

叶宝娟, 温忠麟. (2011). 单维测验合成信度三种区间估计的比较. *心理学报, 43*(4), 453–461.

叶宝娟, 温忠麟. (2012a). 用 Delta 法估计多维测验合成信度的置信区间. *心理科学, 35*(5), 1213–1217.

叶宝娟, 温忠麟. (2012b). 测验同质性系数及其区间估计. *心理学报, 44*(12), 1687–1694.

叶宝娟, 温忠麟. (2013a). α 系数的区间估计方法比较. *心理科学, 36*(1), 215–222.

叶宝娟, 温忠麟. (2013b). 两水平研究中单维测验信度的估计. *心理科学, 36*(3), 728–733.

叶宝娟, 温忠麟, 陈启山. (2012). 追踪研究中测验信度的估计. *心理科学进展, 20*(3), 467–474.

叶宝娟, 温忠麟, 胡竹菁. (2013). 单维测验合成信度元分析. *心理科学, 36*(6), 1464–1469.

叶宝娟, 杨强. (2011). 用验证性因子分析估计单维测验的信度. *教育测量与评价(理论版), 4*(11), 8–12.

叶宝娟, 杨强. (2014). 偏态分布下多维测验合成信度区间估计的比较. *教育测量与评价(理论版)*, *7*(11), 8–11.

叶宝娟, 杨强. (2015). 用 Delta 法估计误差相关测验合成信度的置信区间: 以 FAD 为例. *心理学探新, 35*(3), 251–256.

张力为. (2002). 信度的正用与误用. *北京体育大学学报, 25*(3), 348–350.

张龙飞, 刘凯, 宋鸽, 涂冬波. (2020). 计算机化自适应测验技术在情绪智力智能测评中的初步应用——基于项目反应理论. *江西师范大学学报(自然科学版)*, *44*(5), 454–461.

Alonso, A., Laenen, A., Molenberghs, G., Helena Geys, H., & Vangeneugden, T. (2010). A unified approach to multi-item reliability. *Biometrics, 66*(4), 1061–1068.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74*(1), 137–143.

Edwards, A. A., Joyner, K. J., & Schatschneider, C. (2021). A simulation study on the performance of different reliability estimation methods. *Educational and Psychological Measurement, 81*(6), 1089–1117.

Fu, Y., Wen, Z., & Wang, Y. (2018). The total score with maximal reliability and maximal criterion validity: An illustration using a career satisfaction measure. *Educational and Psychological Measurement, 78*(6), 1108–1122.

Fu, Y., Wen, Z., & Wang, Y. (2022). A comparison of reliability estimation based on confirmatory factor analysis and exploratory structural equation models. *Educational and Psychological Measurement, 82*(2), 205–224.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement, 66*(6), 930–944.

Kelley, K., & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for composite measures. *Psychological Methods, 21*(1), 69–92.

Lai, M. H. C. (2020). Composite reliability of multilevel data: It's about observed scores and construct meanings. *Psychological Methods, 26*(1), 90–102.

Lord, F. M., Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Maydeu-Olivares, A., Coffman, D. L., & Hartmann, W. M. (2007). Asymptotically distribution free (ADF) interval estimation of coefficient alpha. *Psychological Methods, 12*(2), 157–176.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433.

Padilla, M. A., & Divers, J. (2016). A comparison of composite reliability estimators: Coefficient omega confidence intervals in the current literature. *Educational and Psychological Measurement, 76*(3), 436–453.

Pfadt, J. M., van den Bergh, D., Sijtsma, K., Moshagen, M., & Wagenmakers, E. (in press). Bayesian estimation of single-test reliability coefficients. *Multivariate Behavioral Research*.

Raykov, T. (2001). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology, 54*(2), 315–323.

Raykov, T., & Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you! *Educational and Psychological Measurement, 79*(1), 200–210.

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling, 9*(2), 195–212.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. Multivariate Behavioral Research, *47*(5), 667–696.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74*(1), 145–154.

Scherer, R., & Teo, T. (2020). A tutorial on the meta-analytic structural equation modeling of reliability coefficients. *Psychological Methods, 25*(6), 747–775.

Sijtsma, K., & Pfadt, J. M. (2021). Part Ⅱ: On the use, the misuse, and the very limited usefulness of cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrika, 86*(4), 843–860.

ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (in press). Interrater reliability for multilevel data: A generalizability theory approach. *Psychological Methods*.

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for $\omega_h$. *Applied Psychological Measurement, 30*(2), 121–144.

## Research on test reliability in China's mainland from 2001 to 2020

WEN Zhonglin[1], CHEN Hongxi[1], FANG Jie[2], YE Baojuan[3], CAI Baozhen[1]

([1] *School of Psychology & Center for Studies of Psychological Application, South China Normal University, Guangzhou* 510631, *China*) ([2] *Institute of New Development & Department of Applied Psychology, Guangdong University of Finance & Economics, Guangzhou* 510320, *China*) ([3] *School of Psychology & Center of Mental Health Education and Research, Jiangxi Normal University, Nanchang* 330022, *China*)

**Abstract:** With the application of confirmatory factor analysis, research on reliability has entered a new stage. In the first two decades of the 21st century, the studies on test reliability in China's mainland show three main lines of development. The first is the development of test reliability based on confirmatory factor models, including homogeneity coefficient, composite reliability, maximum reliability, etc. The second is the expansion of data types collected by scales, including the reliability of two-level data and longitudinal study. The third is the extended use of reliability, involving rater reliability, encoder reliability, etc. For a common test (with item-errors uncorrelated each other), if the coefficient α is high enough, test reliability is acceptable; otherwise composite reliability is recommended. If the composite reliability of every variable in a statistical model is very high (over 0.95), modeling with composite scores does not differ much from modeling with latent variables. Otherwise, it is better to use latent variable modeling.

**Key words:** reliability, coefficient α, homogeneity coefficient, composite reliability, interval estimation